ESTIMATING EFFECTS OF NON-NORMALITY IN ASSESSING STRUCTURAL EQUATION MODEL FIT FOR USE OF PHYSICAL SCIENCE DATA

by

SARAH ALTA ROSE

DISSERTATION

Submitted to the Graduate School

of Wayne State University

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2016

MAJOR: EDUCATION (Evaluation and Research)

Approved By:

_____

Advisor Date

_____

_____

_____

_____

DEDICATION

I would like to dedicate this dissertation to my family who supported and encouraged me through this, as in every endeavor. My grandmother, Fannie Rose, passed away shortly before the completion of this work; her love and support are truly missed. My grandmother, Marian Frankel, loved education and inspired me to always aim higher academically. My grandfathers, Sam Rose and Leonard Frankel, always gave me their love and support. My father, Mitchell Rose, never ceased to remind me that a formal education doesn't stop until after the doctorate. My mother never ceased to remind me that completing my dissertation was very important. My stepfather has always been an inspiration for me in the continuation of my academic career. My sister, Chaya Thav, reminded me to hold my breath during the difficult times, and that school would be done before I knew it. To my stepmother, Svetlana Rose, and the rest of my brothers and sisters.

ACKNOWLEDGMENTS

I am very grateful to my major advisor, Dr. Barry Markman, who led me through this dissertation process. Thank you for all your help and support. I am also grateful to the other members of my committee, Drs. Monte Piliawsky, Boris Shulkin, Jamie Gleason, and Kraig Warnemuende, who had an important role in the dissertation process. I would like to thank Drs. Shlomo Sawilowsky, Robert Partridge, Bruce Zumbo, and Ryoungsun Park for their consultation and advice concerning the results of the Monte Carlo analyses.

A special thanks to Dr. Shlomo Sawilowsky who had an important role in my educational and professional development.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

Background of the Study

Structural Equation Modeling (SEM) is a relatively new statistical methodology that is beginning to be established in the professional field of statistics. Its foundational theory was published by Wright (1918), where a path analysis was used to model the bone size of rabbits. However, the novelty of the methodology was such that SEM was not accepted by researchers until the 1960s or 1970s (Matsueda, 2011). This coincided with increasing use of computers, allowing for the more practical use of the complicated matrix models by standard researchers.

The development of more complicated analytical procedures was inevitable. Hoyle (1995) indicated, "with the increasing complexity and specificity of research questions in the social and behavioral sciences…has come increasing interest in SEM as a standard approach to testing research hypotheses" (p. 1). Indeed, with the complex nature of many modern research models, it is imperative to use a data analysis tool that allows the most flexibility in the analysis to confirm a best interpretation of the model results. SEM is a powerful tool that can be used to explore data for the purpose of improving the understanding of the interactions, reliability, and general characteristics. It allows for a more complete and comprehensive analysis compared to other research methodologies (such as multiple regression) because it allows freedom in the evaluation of several model construct relationships simultaneously (Alavifar, 2012). This advantage should not be underestimated. The ability to take 5, 10, 20, or 100 variables and analyze them together using one test

without the necessity for Bonferonni or similar corrections allows for considerable increase in statistical power.

SEM has the unique capability to model relationships between variables and to estimate error. SEM can therefore be considered as rather a “union” (p. 3) between path analysis and factor analysis (Gullen, 2000). Modelling error in SEM is a unique advantage. By virtue that error is explicit in the SEM model, as opposed to the implicit implication of error via other methodologies, using SEM can result in a more realistic, reliable, and comprehensive model of the data.

SEM models are developed by determining relationships between observed and/or latent variables to develop an initial model. The model is analyzed to determine whether it is an appropriate approximation of the data construct. If the model is concluded to be an appropriate approximation, it is analyzed to ascertain the magnitude and direction of relationships between the different variables.

Kline (2011) set forth six steps to developing a SEM model. They are (1) specifying the model, (2) evaluating the model identification, (3) selecting the measures and collecting and screening the data, (4) estimating the model, (5) re-specifying the model, and (6) reporting the results.

Probably, the most essential step from the six mentioned above is Step 4. This includes assessing the model to determine how well it represents the data. Many SEM models can be developed that represent the data to a degree; however, a good model will be the best fit representation. To this end, model fit statistics were developed. These statistics result in a quantitative analysis of model fit that allows researchers to determine how well the model fits the data in an objective manner.

The matter of how to develop these fit statistics and which are the best to use has been a topic of great discussion. Kline (2011) indicated that “For at least 30 years the literature has carried an ongoing discussion about the best ways to test hypotheses and assess model fit” (p. 190). There are dozens of fit indices, and each one is a measure of appropriate model fit to the data. For example, IBM’s SPSS Amos Graphics (version 22) provides 20 different fit indices; however, there are dozens of different fit indices that can be considered (Kline, 2011). Most researchers agree that the Chi-Squared test (or Cmin as indicated Amos Graphics) is a basic evaluation of model fit for SEM (Kline, 2011 and Hoyle, 1995) and should be evaluated first. If this test indicates a bad fit, it should weigh considerably on the researcher’s assessment of the model fit.

Other common fit indices include the Root Mean Square Error of Approximation (RMSEA), the Standardized Root Mean Square Residual (SRMR), and the Comparative Fit Index (CFI). These are common fit indices that were recommended by Kline (2011), Hoyle (1995), Byrne (1994), and Hooper, Coughlan, and Mullen (2008). These fit indices are provided by Amos Graphics and are therefore easily obtained. They are discussed further in Chapter 2.

However, other fit indices do exist. These include the Goodness-of-Fit Index (GFI), Adjusted Goodness of Fit Statistic (AGFI), Root Mean Square Residual (RMR), and others (Hooper, Coughlan, & Mullen, 2008).

Each fit index is unique and measures model fit in different manners. For example, the Chi-Square test is based on the “magnitude of discrepancy” (p. 53) between the expected data and the actual data (Hooper, Coughlan, & Mullen 2008).

This test is based on the overall model fit, as opposed to the incremental fit (as will be discussed below). Although the Chi-Square test can be used to assess any model fit distribution; in SEM, the Chi-Square test generally is used to determine variance of the data from normality. The Chi-Squared test has several limitations that affect the Chi-Square values and can provide erroneous approximation of fit. Factors that can inflate or deflate the Chi-Square values include high correlation among observed variables, unique variances among variables, and large samples sizes (Kline, 2011). Additionally, the Chi-Square test gives little information as to the extent that model does not fit (Byrne, 1994). As such, additional statistical measures are necessary to determine model fit approximation of the SEM.

The fit indices for SEM have different limitations and boundary conditions. The necessity for numerous fit indices can be explained in two ways. Firstly, fit indices are greatly important in the performance of any SEM. SEM that is an improper fit to the data would provide inaccurate or erroneous results, and possibly indicate relationships that do not exist. Secondly, the process to performing SEM, the complexity of the variable matrixes and the sheer volume of analysis required, indicate a necessity for numerous fit index models. As the process is rigorous and complicated, so too the fit indexes are difficult to simplify. There is currently no single fit index that encompasses all the different indices in one comprehensive test. (Gullen, 2000)

The complexity of analyzing the fit indices and the plethora of index tests from which to form a model fit assumption, make it necessary to determine when models are truly a good fit to the data.

> Hooper, Coughlan, and Mullen (2008) indicated:
>
> Given the plethora of fit indices, it becomes a temptation to choose those fit indices that indicate the best fit…This should be avoided at all costs as it is essentially sweeping important information under the carpet. (p. 56)

Hooper, Coughlan, and Mullen (2008) recommended several common fit indices to be considered, among which are listed the CFI, the RMSEA and the Chi-Square tests as the most common. However, the tests listed above were developed primarily for social and behavioral science research where the baseline assumption for distribution is normal. It is questionable whether these variables provide a good indication of model fit under different distributions that are common in the physical sciences (i.e. exponential, logarithmic, or uniform).

Problem Statement

The purpose of this study is to evaluate the sensitivity of selected fit index statistics in determining model fit when the distribution varies from normality, as is typically true of data research for the physical sciences. SEM is already being applied to many physical science research problems and the reliability and power of the model fit indices is questionable. Gullen (2000) discussed how “Non-normal data pose problems in structural equation models even if the data are continuous” (p. 19). Gullen (2000), however, did not consider the extent to which the problem is imposed, or which distributions perform better than others. The extent and power of the fit indices in estimating the SEM model fit when normality is violated is therefore of interest.

Theoretical mathematical and physical science distributions and applied physical science data sets will be obtained. Tests for normality will be conducted on the applied data sets. These data sets will be sufficiently large to serve as proxies for typical physical science distributions. The theoretical distributions and applied data sets will then be randomly sampled and subjected to RMSEA, SRMR, and CFI model fit index tests.

Assumption

1. Data sets from real variables and hypothetical distributions will be sampled. It is assumed they are representative of common conditions.

Limitations

1. There are dozens of different fit indices; however, only three common fit indices (RMSEA, SRMR, and CFI) will be used in this study.
2. There are limitless distributions possible. In this study, only three or four variables encompassing different distributions from several real and hypothesized data sets will be analyzed.
3. The results of the fit indices will vary based on the sample size and whether the sample size is balanced or unbalanced between variables. This study will be limited to three or four sample size groups and to balanced sample sizes between variables.
4. The results of the fit indices will vary based on the chosen alpha level. In this study, only one alpha level (of 0.05) will be used.

Definition of Terms

1. Observed Variables: Variables consisting of data that are measurable and provided directly from an instrument.
2. Latent Variables: Variables that consist of a combination of several observed variables, similar to a construct.
3. Model Specification: The process of arranging the observed and latent variables into an SEM model, and specifying their relationships (correlations, errors, and direction).
4. Model Identification: The process of identifying the degrees of freedom of an SEM model.
5. Correlation: The relationship or association between two variables. The correlation can be quantified using the correlation coefficient.
6. Error: Deviation of the data from expected value that cannot be attributed to a model or any substantial explanation.

CHAPTER 2 REVIEW OF LITERATURE

Overview of SEM

SEM is a comprehensive data analysis technique that allows researchers greater flexibility in determining the magnitude and direction of relationships between variables. A primary advantage of SEM lies in the creation of a model. These models can graphically provide quantitative values of variable relationships, correlations, and even error. Other advantages were offered by Chin (1998). Benefits of SEM over "first-generation techniques" (p. vii) (such as factor analysis, principal components analysis, multiple regression, and discriminant analysis) are listed below:

1. Ability to model relationships between variables
2. Ability to model error in observed variables
3. Ability to conduct a priori tests against empirical data (such as in Confirmatory Factor Analysis)
4. Hesketh, Skrondal, & Pickles (2004) offered a fourth advantage in that SEM can be used in the development of latent variables of "hypothetical construct" (p. 168).

SEM can be considered as a conglomeration of several common "first-generation" (Chin, 1998, p. vii) statistical approaches. In running SEM, statistical approaches that are simulated include correlation, multivariate regression, path analysis, maximum likelihood, generalized least squares, and factor analysis. It is for this reason that the SEM methodology contains many similarities with the conventional analytical procedures.

These similarities include:

1. Both are general linear models (Kline, 2011).
2. Analyses are only valid if boundary conditions are met (Weston & Gore, 2006)
3. The techniques do not imply causality (Weston & Gore, 2006)
4. Researchers can misuse the SEM just as with other, more classical, analytical procedures (Weston & Gore, 2006).

Weston and Gore (2006) paradoxically stated:

> "Just as researchers are free (although not encouraged) to conduct several different multiple regression models until they find a model to their liking, they can also analyze models in SEM, identify and remove weaknesses in the model, and then present the revised model as if it were the originally hypothesized model." (p. 723)

One common way that researchers can misuse models includes publicizing the fit index measures that indicate a good model fit and neglecting the indices that indicate a poor fit. As indicated above, there are a plethora of model fit indices and each one measures model fit in a different way. Additionally, as the distribution of the samples vary from the boundary conditions, as set forth by the SEM computer software (i.e. normality); the fit indices can result in erroneous assessments that can inflate model fit statistics. It is therefore of importance to understand how variations from normality affect the model fit equations.

Kline's (2011) Six Steps to Performing SEM

SEM is a relatively new statistical procedure that is tailored for a rigorous analytical approach of data research. Even using a computer program to solve for the matrix algebra and determine the output results, the approach for analyzing the results and for selecting the measures is complex. Kline (2011) simplified the procedures by identifying six steps to performing SEM. These steps, stated in Chapter 1, are explained in fuller detail below.

Step 1 - Specifying the Model: Specifying the model includes analysis of the relationships between the variables and drawing a model diagram. Model diagrams include endogenous and exogenous variables that are typically represented by arrows, indicating direction. Correlations are typically represented by a circular, two-direction arrow, indicating a strong relationship between two variables. SEM models assume that variables without correlations are not highly correlated. Ideal SEM models include primary variables that are moderately correlated with coefficient values ranging between 0.4 and 0.7.

Step 2 - Model Identification: Model identification is necessary prior to beginning model estimation to determine whether it is "*theoretically* possible for the computer to derive a unique estimate of every model parameter." (Kline, 2011, p. 93). An under-identified model cannot provide reliable model results. Therefore, prior to model estimation, it is imperative to determine that the model is over-identified or just-identified. Gullen (2000) compared this to algebra, as numerous equations and variables are required. "Unlike in algebra, however, there is a benefit to having more equations than variables. These over-identified models permit the calculation of fit

statistics for the evaluation of model fit." (p. 1). In other words, just-identified models can allow for a unique solution but the fit indices would not provide valid results for determination of model fit (see Step 4).

Under-identified, just-identified, and over-identified designations are a function of the model complexity. SEM analyses generally provide improved results with a more parsimonious model. A parsimonious model is a model with a larger degree of freedom. The degrees of freedom can be calculated per Kline (2011) using the following equation:

$$df_m = p - q \quad \text{Eq. (1)}$$

$$p = v(v+1)/2\ , \quad \text{Eq. (2)}$$

where v = number of observed variables, and q = number of estimated parameters.

A model with fewer estimated parameters (i.e. correlations, error terms, etc.) and many observed variables would be more parsimonious. Parsimony is preferable, and many model fit indices include an adjustment to account for parsimonious models (refer to the subsection "Approximate Fit Indices" below).

Step 3 – Selecting the Measures: Selecting the measures and collecting/preparing data involve determining whether the data are appropriate for an SEM model. This includes ensuring that extreme collinearity does not occur. Extreme collinearity occurs when the primarily (latent) variables are highly correlated with each other. This step also includes dealing with outliers and missing data as prescribed by standard statistical procedures.

Step 4 – Model Estimation: Model estimation means determining whether the SEM is a good estimation of the data. This involves determining model fit based on